DOCUMENT RESUME

ED 424 300 TM 029 192

AUTHOR Webster, William J.; Mendro, Robert L.; Orsak, Timothy H.; Weerasinghe, Dash

TITLE An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect.

PUB DATE 1998-04-00

NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Information Analyses (070) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Elementary Secondary Education; *Estimation (Mathematics); Mathematical Models; *School Effectiveness; *Teacher Effectiveness

IDENTIFIERS *Criterion Variables; *Hierarchical Linear Modeling

ABSTRACT

This paper provides a concise summary of 10 years of research into defining appropriate statistical models for estimating school and teacher effect on student learning and other important educational outcomes. It discusses criteria for judging models and presents the formulae for the two-stage, two-level student-school hierarchical linear modeling (HLM) model that is the model of choice for estimating school effect and the two-stage, two-level student-teacher HLM model that is the model of choice for estimating teacher effect. Finally, a brief discussion is provided on criterion variables and the methods by which they are weighted, predicted, and aggregated in the school and teacher effectiveness system. (Contains 1 figure and 39 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
William Webster
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

An Application of Hierarchical Linear Modeling to the Estimation of School and Teacher Effect[1]

William J. Webster, Robert L. Mendro,
Timothy H. Orsak, and Dash Weerasinghe

Dallas Public Schools

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

This paper provides a concise summary of ten years of research into defining appropriate statistical models for estimating school and teacher effect on student learning and other important educational outcomes. It discusses criteria for judging models and presents the formulae for the two-stage, two-level student-school HLM model that is the model of choice for estimating school effect and the two-stage, two-level student-teacher HLM model that is the model of choice for estimating teacher effect. Finally, a brief discussion is provided on criterion variables and the methods by which they are weighted, predicted, and aggregated in the school and teacher effectiveness system.

The need for instructional improvement in the Dallas Public Schools had been thoroughly documented over a period of twenty years. After a period of rapid achievement growth in the early and mid 1980's, student achievement in the Dallas schools had leveled out. In 1990, responding to this need, the District's Board of Education appointed a citizen's task force, the Commission on Educational Excellence, to formulate recommendations to accelerate the needed improvement. After a year of community hearings and extensive study, the Commission recommended a six point plan for massive educational reform. At the heart of the Commission's recommendations was an accountability system that fairly and accurately evaluated schools and teachers on their contributions to accelerating student growth in a number of important and valued outcomes schooling. This was coupled with a movement to give schools more decision-making authority about personnel, curriculum, and most other aspects of schooling. In exchange for this authority, school staffs were to be held accountable for their actions.

[1] Paper presented at the annual meeting of the American Educational Research Association, April 13-17, 1998.

BEST COPY AVAILABLE

ED 424 300

TM029192

As part of this recommendation, $2.4 million was set aside as an incentive award to reward effective schools, and their professional and support staffs.

It then became the task of the District's Research, Planning, and Evaluation Department to develop, pilot test, and implement an evaluation system to accomplish the goals of the Commission. The first step in accomplishing this task was the appointment of an Accountability Task Force to oversee the process. This task force, consisting of teachers, principals, parents, members of the business community, and central office administrators, was charged with the responsibility of advising the General Superintendent concerning the implementation of a performance incentive plan, working with the administration to ensure the validity of the selection procedure and subsequent results of the incentive plan, and serving as a review committee to examine any issues raised by personnel concerning questions of equity and fairness of the procedures. During a year of exhaustive deliberations, a number of requirements for the methodology associated with this plan were developed. Among these were:

1. It must be value-added.

2. It must include multiple outcome variables.[2]

3. Schools must only be held accountable for students who have been exposed to their instructional program (continuously enrolled students).

4. It must be fair. Schools must derive no particular advantage by starting with high-scoring or low-scoring students, minority or white students, high or low socioeconomic level students, or limited English proficient or non-limited

---

[2] Performance indicators for 1997-1998 include *Iowa Tests of Basic Skills* and *Tests of Achievement and Proficiency* reading and mathematics, grades 1-9; *Spanish Assessment of Basic Education*, grades 1-6; *Texas Assessment of Academic Skills*, reading and mathematics, grades 3-8 and 10; writing, grades 4, 8, and 10; science and social studies, grades 4 and 8; *Texas Assessment of Academic Skills*, Spanish version grades 3-6; 143 standardized final examinations in language, mathematics, social studies, science, ESOL, reading, and world languages, grades 9-12; promotion rate, grades 1-8; student attendance, grades 1-12; graduation rate, grades 9-12; *Scholastic Aptitude Test* percent tested and scores, grades 9-12; dropout rate, grades 7-12; student enrollment in prehonors/honors courses, grades 7-12; student enrollment in advanced diploma plans, grades 9-12; students enrolled in advanced placement courses, grades 11-12; *Preliminary Scholastic Aptitude Test* percent tested and scores, grades 9-12; and percent passing *Advance Placement Exams*, grades 11-12. The system is operated with only continuously enrolled students and includes staff attendance incentives, minimum percent eligible tested requirements, and requirements that at least one-half of a school's cohorts must outgrow the national norm group on the *ITBS* and *TAP* in reading and mathematics.

English proficient students. In addition such factors as student mobility, school overcrowding, and staffing patterns over which the schools have no control must be taken into consideration.

5. It must be based on cohorts of students, not cross-sectional data.

Within the five aforementioned parameters, a number of statistical models are possible. This study examines alternative methodologies for determining school effect then extends these studies to the determination of teacher effect. These models are designed to isolate the effect of a school's or teacher's practices on important student outcomes. The school effect can be conceptualized as the difference between a given student's performance in a particular school and the performance that would have been expected if that student had attended a school with similar context but with practice of average effectiveness. The teacher effect can be conceptualized similarly at the teacher level.

Background

Interest in performance-based or outcomes-based teacher evaluation dates all the way back to fifteenth-century Italy where a teacher master's salary was dependent upon his or her students' performance. Despite long-term interest, progress in actually linking student outcomes to school and teacher performance has been very limited.

State Departments of Education have taken a leadership role in attempting to accomplish this at the school and district level. Forty-six of fifty states have accountability systems that feature some type of assessment. Twenty-seven of these systems feature reports at the school, district, and state level, three feature school level reports only, six feature reports at both the school and district level, seven feature reports at the district and state level, two feature reports at the state level only, and one is currently under development (Council of Chief State School Officers, 1995). When one reviews these systems, it is obvious that their designers are not familiar with the literature on value-added systems since only two states, South Carolina (May, 1990) and Tennessee

(Sanders and Horn, 1995) have used appropriate value-added statistical methodology in implementing such systems. Most of the rest tend to evaluate students, not schools or districts, and generally cause more harm than good with systematic misinformation about the contributions of schools and districts to student academic accomplishments. By comparing schools on the basis of unadjusted student outcomes, state reports are often systematically biased against schools with population demographics that differ from the norm, a fact that was graphically illustrated by Jaeger (1992). In attempting to eliminate this bias, a number of states have gone to non-statistical grouping techniques, an approach that has serious limitations when there is consistent one-directional variance on the grouping characteristics within groups.

Investigators throughout the world have conducted and reported numerous studies aimed at identifying effective schools as well as estimating the magnitude and stability of school contributions to student outcomes. Good and Brophy (1986) provide an excellent review of this work. Researchers have been working for a number of years on appropriate methodology for adjusting for the effects of student and school demographic variables in estimating school effects. One approach has been to regress school mean outcome measures on school means of one or more background variables. This approach is only adequate to the extent that there is not much within school variance, that is, the school impacts all students similarly. Mendro and Webster (1993) demonstrated that this is generally not the case and that using school level models to attempt to estimate school effects, while better than the common practice of reporting unadjusted test scores, produces extremely unstable estimates of school effects.

Another approach, one that has received generally widespread acceptance among educational researchers, involves the aggregation of residuals from student-level regression models (Aiken and West, 1991; Bano, 1985; Felter and Carlson, 1985; Kirst, 1986; Klitgard and Hall, 1973; McKenzie, 1983; Millman, 1981; Saka, 1984; Webster and Olson, 1988; Webster, Mendro, and Almaguer, 1994). These techniques can incorporate a large number of input, process, and outcome variables into an equation and

determine the average deviation from the predicted student outcome values for each school. Schools are then ranked on the average deviation. Some advantages of multiple regression analysis over other statistical techniques for this application include its relative simplicity of application and interpretation, its robustness, and the fact that general methods of structuring complex regression equations to include combinations of categorical and continuous variables and their interactions are relatively straightforward (Aiken and West, 1991; Cohen, 1968; Cohen and Cohen, 1975; Darlington, 1990).

Finally, hierarchical linear modeling (HLM) provides estimates of linear equations that explain outcomes for group members as a function of the characteristics of the group as well as the characteristics of the members. Because HLM involves the prediction of outcomes of members who are nested within groups which in turn may be nested in larger groups, the technique should be well suited for use in education. The nested structure of students within classrooms and classrooms within schools produces a different variance at each level for factors measured at that level. Bryk, et. al. (1988) cited four advantages of HLM over regular linear models. First, it can explain achievement and growth as a function of school level or classroom level characteristics while taking into account the variance of student outcomes within schools or classrooms. Second, it can model the effects of student characteristics, such as gender, race/ethnicity, or socioeconomic status, on achievement within schools or classrooms and then explain differences in these effects between schools or classrooms using school or classroom characteristics. Third, it can model the between and within group variance at the same time and thus produce more accurate estimates of student outcomes. Finally, it can produce better estimates of the predictors of student outcomes within schools and classrooms by using information about these relationships from other schools and classrooms. HLM models are discussed in the literature under a number of different titles by different authors from a number of diverse disciplines (Bryk and Raudenbush, 1992: Dempster, Rubin and Tsutakawa, 1981; Elston and Grizzle, 1962; Goldstein, 1987; Henderson, 1984; Laird and Ware, 1982; Longford, 1987; Mason, Wong, and Entwistle, 1984; Rosenburg, 1973).

Extending this methodology to the teacher level becomes even more complex. The issue really is not one of whether or not student achievement data should be used in teacher evaluation, but rather entails a methodological debate over ways to operationalize and implement such a system. Unfortunately, the preponderance of literature in the field concentrates upon reasons student achievement data cannot be used for teacher evaluation rather than upon credible ways to use it. Some of the concerns raised in the literature include:

- the development of procedures to account for the difficulty in measuring the long-term development of skills which may not be measured in year-to-year growth patterns (TEA, 1988).

- the assessment of diverse areas of achievement which do not have readily available standardized tests is an area of concern when dealing with non-academic area teachers.

- programs which pull out students for remediation, programs which involve team-teaching, and programs with extensive use of instructional aides inhibit the estimation of an individual teacher's contribution to improved student achievement.

- norm-referenced standardized tests sample broad subject domains and are unlikely to match closely the curriculum in particular classrooms at particular times (Haertel, 1986).

- well-established, broadly applicable, and accepted achievement measures are not available in all the relevant areas of learning (Bano, 1985).

- standardized achievement tests are unlikely to reflect the full range of instructional goals in their subject areas. Norm-referenced tests tend to ignore the higher-order skills. Therefore it is likely that products of superior teaching are not measured adequately or completely by standardized achievement tests (Bano, 1985).

- what the student brings to the classroom in terms of ability, home and peer influence, motivation and other influences is very powerful in affecting academic achievement at the end of the year (Iwanicki, 1986).

- the statistical methods used to control for non-teacher factors cannot take into account all of the relevant factors. More importantly, the methods will be

incomprehensible to those being evaluated and difficult to defend in public (Bano, 1985).

- non-statistical models for controlling non-teacher factors are easier to explain, but cannot take into account most of the necessary circumstances (Bano, 1985).

- attempting to use any one of a number of regression-based techniques at the teacher level creates a rather subtle problem related to the statistical concept of "degrees of freedom." In general, the number of degrees of freedom upon which a statistical procedure is based depends on the sample size (N) and the number of sample statistics (i.e., variables in multiple regression). The sample size (i.e., number of students) for a teacher is relatively small to start with. However, the usable sample size becomes even smaller because development of the regression equations requires existing test scores for each student for at least two successive years. As an example, a second-grade teacher may have a class of 22 students, but may only have test scores from the first grade for 11 of those students. Since degrees of freedom also depends on the number of variables in the multiple regression equation, a regression equation with four (4) variables would leave just seven (7) degrees of freedom. The stability of a projected regression line is primarily dependent on the number of degrees of freedom. Seven is generally not enough for stable estimates. As a general rule of thumb, thirty students per variable has been recommended as a minimum number upon which to base a projected regression line.

Nontechnical concerns most often found in the literature include the concern that objectives that are not measured by the tests will be omitted by teachers, that other duties such as playground supervision and school committee work may be slighted, and that, with each teacher being rated separately, the collegiality necessary to building good instructional teams within a building may be damaged.

Most of the methodological issues raised above can be resolved. (1) Longitudinal growth curves, or alternatively, relationships based upon two years of data, can be formulated on important outcome variables. In the case of relationships based upon two years of data, replication is necessary to assure greater reliability. (2) Criterion-referenced tests can be developed and used to assess diverse areas of achievement. (3) In cases where there are pull-out or send-in programs, team teaching, or instructional aides, data can be provided at the team level rather than at the individual teacher level. (4) Measures in addition to norm-referenced tests can be used. (5) Constituents are primarily

interested in basic skills. To the extent that measures are needed in music, art, physical education, etc., they can be developed. (6) Criterion-referenced tests can be used to measure higher-order thinking skills. In addition, performance testing can be used as one outcome variable with the outcomes being weighted by the reliability of the instruments. (7) What the student brings to the classroom in terms of background variables can be statistically controlled. These variables typically account for 9-20% of the variance in student achievement (Webster, Mendro, and Almaguer, 1993). (8) It has been the author's experience that gender, ethnicity, limited English proficiency status, and free-or-reduced-lunch status, plus their interactions, account for most of the variance that can be attributed to background variables. They are easy to explain or defend. (9) Non-statistical models for controlling non-teacher factors are misleading and should not be used (Webster and Edwards, 1993). (10) The degrees of freedom problem is real in that one must worry about the stability of the regression line when it is applied to one teacher. At the teacher level, replication over several years is the best safeguard against erring because of small sample size.

Criteria For Judging Models

The traditional criteria for judging the efficacy of regression-based models is goodness of fit (r^2). This is not a particularly useful tool for judging the effects of statistical models that are designed to rank schools or teachers. If one uses the individual student as the unit of analysis and applies either OLS regression or HLM to estimating effect, the differences in r^2s produced by these models is minute.

The criterion that we believe should be used in judging the appropriateness of statistical models designed to rank schools and teachers is fairness. That is, a school's or teacher's estimated effectiveness level should not be capable of being predicted by the individual or aggregate composition of its student body or classrooms. Variables over which the school or teacher have no control should not be correlated with the school's or teacher's effectiveness rating. These variables include such things as student pretest score, ethnicity, economic status, mobility, gender, and limited English proficient status.

The models that are proposed in this paper produce results at the school level that correlate zero with important school and classroom level contextual variables.

It is obviously also very important that there be a school or teacher effect. If there is none, one is reduced to ranking schools based on random error. This methodology must be part of a comprehensive accountability system that provides valid data for decision-making and improvement as well as for accountability (Webster, 1998).

Relevant Results

All studies summarized in this section examined correlations between indices produced by various methods and those produced by the methodology of choice as well as correlations with individual student background and classroom and school contextual variables. The methodology of choice for producing school effectiveness estimates is a two-stage, two-level student-school HLM model while that for producing estimates of teacher effect is a two-stage, two-level student-teacher HLM model. Formulae for both of these models are specified later in this paper. Results are discussed in relation to the models of choice in order to limit the statistics presented and simplify the discussion. Detailed backup data are contained in the various referenced papers.

School Level

Individual student background variables included gender, socioeconomic status (free or reduced lunch, parental income, family poverty index, parental education level), ethnicity, limited English proficient status, and pretest scores. Aggregate school variables include percent student mobility, percent overcrowded, percent economically disadvantaged (same variables as specified above), percent limited English proficient, percent Black, percent Hispanic, and percent minority. The majority of studies were done at the grades 4-6 levels although additional studies were conducted at grade 8. Conducting studies at the different grade levels is significant because the elementary levels (grades 4-6) have large numbers of relatively homogeneous schools while the middle school level (grade 8) has a small number of relatively heterogeneous schools.

Before discussing a number of thoughtful methodologies for estimating school effect, it is important to reiterate that ranking schools based on unadjusted student test scores or on gain scores is neither particularly informative nor fair. The results produced by these systems correlate poorly with the results produced by the model of choice (r's$\leq$ .508 for unadjusted test scores, r's$\leq$.732 for gain scores) and produce results with unacceptably high correlations with student background variables and school aggregate variables (as high as r=.648 with parental education level) (Webster, et.al., 1995). Evidence suggests that this type of reporting under the guise of determining school or teacher effect does severe injustice to teachers and schools that serve poor and minority student populations. It is important to note that the backbone of most state accountability systems is unadjusted test scores or student gain scores, often not even based on cohort data.

Ordinary least squares regression (OLS) models improve reporting significantly over those models discussed in the previous paragraph as long as the models use data at the individual student level. Using OLS models with aggregated school level variables produces results that correlate poorly with results produced by the model of choice ($r\leq$.58) as well as correlating highly with student and school level contextual variables. Too much information is lost when student data are aggregated to the school level prior to analysis. The greater the within group variance of the individual schools the poorer the estimates produced by the aggregate models (Mendro and Webster, 1993).

OLS models that include all of the individual student demographic variables presented above as well as relevant pretest scores produce results that are moderately correlated with the model of choice ($r\leq$.8637) and are relatively unbiased at the individual student level (most r's $\leq$ .02). Correlations with background variables become higher than desired at the school level with the correlation with percent Black reaching -.1225. If one looks at a grade with relatively few schools (<30) with high within group variance such as grade 8 in this series of studies, the school level correlations explode to $r\geq$.40 for most socioeconomic status indicators (Webster, et.al., 1997).

Implementing a two-stage OLS regression model with student demographic variables regressed on pretest and posttest variables in stage one and the resulting

residuals used in a stage two model predicting posttest from relevant pretests does little to improve the equations. Correlations between the indices produced by the one stage versus two-stage models consistently hover around $r \geq .95$ and the results from the two-stage model correlate slightly higher with the model of choice ($r \leq .8878$). However, correlations with important student background variables are not improved and the correlations with school level contextual variables are actually slightly higher (Webster, et.al., 1995, 1996, 1997). The reason for the development of the original two-stage OLS equations was ease of explanation not statistical parsimony.

The final series of OLS regression equations examined in these series of studies utilized individual student growth curves based on two, three, and four years of data.. No demographic data were included in the equations since it was believed that the individual student growth curves would serve as a surrogate for student background variables. These equations produced results that correlated poorly with the other OLS regression models ($r \leq .85$), more poorly with the model of choice ($r \leq .75$), had unacceptably high correlations with student ethnicity (r as high as -.3587 with Black students), and registered unacceptable correlations with school level contextual variables (r as high as .4621 with percentage Black). Thus it seems obvious that individual student growth curves do not contain all of the information necessary for non-biased prediction. As an aside, the results produced by using four years of prediction correlated .9992 with the results produced using three years of prediction and included about 5% more of the student population (Webster and Olson, 1988; Webster, et.al., 1997).

Moving to hierarchical linear models (HLM), a number of questions were investigated across a series of studies. The first involved whether or not the use of a two or three-level HLM model would produce improved effectiveness indices, that is, indices that were not correlated with student level or school level contextual variables. It is important to note that the correlation between comparable OLS regression models and two-stage, two-level HLM models was $r \geq .97$. (The correlation of the two-stage OLS regression model with the HLM model of choice was only $r \leq .8878$ because the HLM model of choice had additional student background and school contextual variables.) Both models (OLS and HLM) produced minimal correlations with student level

background variables ($r \leq .01$) but the OLS regression model produced correlations with school level contextual variables as high as -.1794 while the HLM model caused all of those correlations to be zero (Webster, et.al., 1995, 1996, 1997).

The basic three-level HLM model that was designed to include comparable student and school-level contextual variables would not run in either a one-stage or two-stage form. Although several models were attempted, major problems were encountered with the algorithms for solving them. In short, in order to successfully run a three-level HLM status model many important contextual variables had to be eliminated from the equations resulting in models that produced unacceptably high correlations with non-controlled contextual variables (Webster, et. al., 1995).

We also examined a three-level HLM model using gain scores as the unit of analysis instead of pretest and posttest scores. We compared the results obtained from comparable two level student-school pretest-posttest HLM models with those obtained from three-level gain score-student-school HLM models and got virtually identical results (r's $\geq .98$). Two level models are more convenient and efficient than three level models because they can accommodate more level one student and level two school contextual variables and they are not nearly as sensitive to multicollinearity and low variance in conditioning variables as are three-level models. Whether fixed or random slopes are assumed, the number of second and third level conditioning variables are severely limited in the three-level model. The inability to accommodate sufficient conditioning variables in the three-level HLM gain score model causes results that correlate poorly with the model of choice ($r \leq .40$) and produces correlations as high as .1887 with important student background variables as well as producing results that are highly correlated with important school contextual variables that the models are not able to accommodate (r's as high as .4747) (Webster, et.al., 1997; Weerasinghe, et.al., 1997).

Throughout the course of the various studies several other important issues were investigated. The issue of one-stage versus two-stage models, be they OLS regression or HLM, is moot. Correlations between and among one-stage versus two-stage models consistently hover around $r \geq .95$ for comparable models and there generally is no practical difference when results are correlated with background variables (Webster,

et.al., 1997). However, the correlations of residuals produced by one-stage HLM models with student level contextual variables suggest that one-stage HLM models carry suppresser effects that are not found in two-stage HLM models or OLS regression models. When this result is coupled with the inability to include important school level contextual variables in one-stage HLM models because of limitations of the models resulting in unsatisfactory correlations with those contextual variables, two stage models are the models of choice.

The final issue investigated was the fixed versus random slopes issue. Correlations between and among comparable models assuming fixed versus random slopes were generally around $r \geq .98$. Models studied were all two-stage HLM models since one-stage HLM models including a full array of contextual variables and assuming random slopes could not be solved. These models produced low correlations with student-level background variables and, when school level conditioning variables were added, zero correlations with school level variables. Since the random model controls for the effects of possible interactions of concomitant variables in specific school settings, and there is no difficulty in solving the two-stage, two-level HLM equations assuming random slopes, our models are random slope models.

Formulae for Estimating School Effect

Based on the analyses conducted through the series of studies reported in this paper, the authors believe that an HLM two-stage, two-level random model with a full range of student and school level contextual variables produces the most bias-free estimates of school effect.

Student level variables included in the HLM models are:

Y_{ij} = Outcome variable of interest for each student *i* in school *j*.
X_{1ij} = Black English Proficient Status (1 if black, 0 otherwise).
X_{2ij} = Hispanic English Proficient Status (1 if Hispanic, 0 otherwise).
X_{3ij} = Limited English Proficient Status (1 if LEP, 0 otherwise).
X_{4ij} = Gender (1 if male, 0 if female).
X_{5ij} = Free or Reduced Lunch Status (1 if subsidized, 0 otherwise).
X_{6ij} = Block Average Family Income.
X_{7ij} = Block Average Family Education.

X_{8ij} = Block Average Family Poverty Level.
X_{kij} = Indicates the variable k of i^{th} student in school j for $i = 1,2, \ldots, I_j$ and $j = 1, 2, \ldots, J$.

School level variables included in the HLM models are:

W_{1j} = School Mobility.
W_{2j} = School Overcrowdedness.
W_{3j} = School Average Family Education.
W_{4j} = School Average Family Education.
W_{5j} = School Average Family Poverty Index.
W_{6j} = School Percentage on Free or Reduced Lunch.
W_{7j} = School Percentage Minority.
W_{8j} = School Percentage Black.
W_{9j} = School Percentage Hispanic.
W_{10j} = School Percentage Limited English Proficient.
W_{11j} = School Percentage Instructional Days Lost to Unfilled Vacancies.

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10} (X_{2ij} X_{4ij}) + \Lambda_{11} (X_{3ij} X_{4ij}) + \Lambda_{12} (X_{1ij} X_{5ij}) + \Lambda_{13} (X_{2ij} X_{5ij}) + \Lambda_{14} (X_{3ij} X_{5ij}) + \Lambda_{15} (X_{4ij} X_{5ij}) + \Lambda_{16} (X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17} (X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18} (X_{3ij} X_{4ij} X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$\text{Criterion Variable_R_98}_{ij} = \beta_{0j} + \beta_{1j} R_1_97_{ij} + \ldots + \beta_{nj} R_n_97_{ij} + \delta_{ij}$$

where

$\text{Criterion Variable_R_98}_{ij}$ = posttest residual from stage one
R_n_97 = n^{th} pretest residual from stage one
$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$.

Level 2:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_{1j} + \gamma_{k2} W_{2j} + \ldots + \gamma_{k11} W_{11j} + u_{kj}$$
for $k = 0, 1, 2, \ldots, n$.

$E[u_{kj}] = 0$, $\text{Var/Cov}[u_{kj}] = T$, and $u_{kj} \perp \delta_{ij}$

$$SEI_j^* = u_{0j}^*$$

Teacher Level

Effectiveness indices at the teacher level are somewhat more complex and require great care in interpretation and use. The Dallas Public Schools uses classroom effectiveness indices as part of the needs assessment in the teacher evaluation system, not as an evaluative tool per se. Teachers are required to formulate strategies to remediate problems detected through the classroom effectiveness indices, student skills analyses, and other formative data. They are then evaluated on how well they implement those strategies.

Since we are under severe time constraints in the production of classroom effectiveness indices, the most parsimonious solution was sought. Working off the information that we had obtained in the school effectiveness indices research, we had hoped to be able to disaggregate the school data to classroom level, apply an adjustment for shrinkage, and produce classroom effectiveness indices. The indices produced in this manner correlate very highly with those produced by other legitimate models, but produce unacceptably high correlations with classroom level contextual variables (Sanders,1997; Webster, et.al., 1997). Thus, although it is more time-consuming to compute, the model of choice for producing classroom effectiveness indices is a two-stage, two-level student-classroom random model HLM. [3]

Formulae for Estimating Teacher Effect

Student level variables included in the HLM classroom effects models are:

Y_{ij} = Outcome variable of interest for each student *i* in classroom *j*.
X_{1ij} = Black English Proficient Status (1 if black, 0 otherwise).
X_{2ij} = Hispanic English Proficient Status (1 if Hispanic, 0 otherwise).
X_{3ij} = Limited English Proficient Status (1 if LEP, 0 otherwise).
X_{4ij} = Gender (1 if male, 0 if female).
X_{5ij} = Free or Reduced Lunch Status (1 if subsidized, 0 otherwise).
X_{6ij} = Block Average Family Income.

[3] Interesting enough, the model that produces the most bias-free estimates at the classroom level without using classroom conditioning variables is a two-stage OLS regression model. Thus, if one is under serious time constraints, a two-stage OLS regression model with an adjustment for shrinkage would produce results that are very close to the model of choice.

X_{7ij} = Block Average Family Education.
X_{8ij} = Block Average Family Poverty Level.
X_{kij} = Indicates the variable k of i^{th} student in classroom j for $i = 1,2, \ldots, I_j$ and $j = 1, 2, \ldots, J$.

Classroom level variables included in the HLM models are:

T_{1j} = Classroom Mobility.
T_{2j} = Classroom Overcrowdedness.
T_{3j} = Classroom Average Family Education.
T_{4j} = Classroom Average Family Education.
T_{5j} = Classroom Average Family Poverty Index.
T_{6j} = Classroom Percentage on Free or Reduced Lunch.
T_{7j} = Classroom Percentage Minority.
T_{8j} = Classroom Percentage Black.
T_{9j} = Classroom Percentage Hispanic.
T_{10j} = Classroom Percentage Limited English Proficient.

STAGE 1:

$$Y_{ij} = \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 (X_{1ij} X_{4ij}) + \Lambda_{10} (X_{2ij} X_{4ij}) + \Lambda_{11} (X_{3ij} X_{4ij}) + \Lambda_{12} (X_{1ij} X_{5ij}) + \Lambda_{13} (X_{2ij} X_{5ij}) + \Lambda_{14} (X_{3ij} X_{5ij}) + \Lambda_{15} (X_{4ij} X_{5ij}) + \Lambda_{16} (X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{17} (X_{2ij} X_{4ij} X_{5ij}) + \Lambda_{18} (X_{3ij} X_{4ij} X_{5ij}) + r_{ij}$$

STAGE 2:

Level 1:

$$\text{Criterion Variable_R_98}_{ij} = \beta_{0j} + \beta_{1j} R_1_97_{ij} + \ldots + \beta_{nj} R_n_97_{ij} + \delta_{ij}$$

where

Criterion Variable_R_98$_{ij}$ = posttest residual from stage one
R_n_97 = n^{th} pretest residual from stage one
$\delta_{ij} \overset{iid}{\sim} N(0,\sigma^2)$.

Level 2:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} T_{1j} + \gamma_{k2} T_{2j} + \ldots + \gamma_{k10} T_{10j} + u_{kj}$$
for $k = 0, 1, 2, \ldots, n$.

$$E[u_{kj}] = 0, \text{Var/Cov}[u_{kj}] = T, \text{ and } u_{kj} \perp \delta_{ij}$$

$$CEI_j^* = u_{0j}^*$$

Outcome Variables and Associated Equations

Figure 1 shows the nature of the equations used in the generation of school effectiveness indices. Each outcome variable is described under "Outcome" along with the grades at which it is included, the score that is the basis for the analysis, the methodology utilized, the level at which the data are analyzed (student or school level), possible predictors and the grades at which they are found, and the school level conditioning variables included in the student level equations. Two different regression models are used depending on whether the unit of analysis is the student, in which case hierarchical linear modeling is used, or the school, in which case multiple regression analysis is used. Through these approaches it is possible to obtain extremely reliable predictions of student and school outcomes and to compare actual outcomes to those that are predicted. All analyses that are done at the student level are calculated on residuals, that is, statistics that have had individual student characteristics over which the schools have no control removed from the equations (gender, ethnicity, limited English proficient status, socioeconomic status, and all of the interactions between those variables).

Classroom effectiveness indices are computed utilizing student and classroom data (classroom level conditioning variables) for the *Iowa Tests of Basic Skills*, the *Tests of Achievement and Proficiency*, the *Texas Assessment of Academic Skills*, the *Texas Assessment of Academic Skills-Spanish*, the *Spanish Assessment of Basic Education*, the *Assessments of Course Performance*, and the *Woodcock-Muñoz Language Survey*.

Summary

This paper has described statistical models for estimating school and teacher effect. It has summarized ten years of research on developing these models and has specified the models in use in the Dallas Public Schools. The school and classroom effectiveness indices described in this paper are part of a comprehensive evaluation system that provides data for both decision-making and accountability. That system is focused on continuous improvement and includes both absolute measures through the *School* and *District Improvement Plans* and relative measures through the school and classroom effectiveness indices. Webster (1998) provides a comprehensive overview of that system.

Figure 1. Description of Variables and Methodology

OUTCOME	GRADES	METHODOLOGY	POSSIBLE PREDICTIONS	GRADES	SCHOOL LEVEL FAIRNESS VARIABLES
1. *Iowa Tests of Basic Skills*, year n, Reading and Mathematics (raw score)	1-8	HLM on residuals (student level)	*Iowa Tests of Basic Skills*, year n-1, Reading and Mathematics	K-7	School mobility, school overcrowdedness, school level average family income, school level average family education level, school level average poverty index, school level percent students on frec/reduced lunch, school level percent limited English proficient students, school level percent Black, Hispanic, and minority students, school level percent instructional days lost to medical disability leave and unfilled vacancies.
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	3-7	
2. *Tests of Achievement and Proficiency*, year n, Reading and Mathematics (raw score)	9	HLM on residuals (student level)	*Iowa Tests of Basic Skills*, year n-1, Reading and Mathematics	8	Same as #1.
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	8	
			Tests of Achievement and Proficiency, year n-1, Reading and Mathematics score (for students with insufficient credits to be in tenth grade)	9	
3. Promotion Rate, year n (percent promoted)	1-6, 7-8	Multiple regression (school level)	Promotion rate in years n-1 and n-2.	1-6, 7-8	None

BEST COPY AVAILABLE

4. Student Attendance, year n (days attended)	1-12	HLM on residuals (student level)	Student Attendance, year n-1	K-11	Same as #1.
5. *Texas Assessment of Academic Skills*, year n, Reading and Mathematics (raw score)	3-8, 10	HLM on residuals (student level)	Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	3-7	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	2-7	
			Tests of Achievement and Proficiency, year n-1, Reading and Mathematics	9	
6. *Texas Assessment of Academic Skills* Spanish, year n, Reading and Mathematics (raw score)	3,4,5,6	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	2,3,4,5	Same as #1
7. *Texas Assessment of Academic Skills*, year n, Writing (raw score) (Spanish writing at grade 4)	4, 8, 10	HLM on residuals (student level)	*Texas Assessment of Academic Skills*, year n-1, Reading and Mathematics	3,7,9	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	3,7	
			Tests of Achievement and Proficiency, Reading and Mathematics	9	
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	3,7,9	
8. *Texas Assessment of Academic Skills*, year n, Science and Social Studies (raw score)	8	HLM on residuals (student level)	*Texas Assessment of Academic Skills*, year n-1, Reading and Mathematics	7	Same as #1.
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	7	

ST COPY AVAILABLE

9. *Spanish Assessment of Basic Education*, reading and mathematics, year n (raw score)	1-6	HLM on residuals (student level)	*Spanish Assessment of Basic Education*, year n-1, Reading and Mathematics	1-5	Same as #1.
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	K-5	
10. Appropriate ESOL ACP, year n (There are five ESOL ACPS, ESOL 1-3, Reading, and ESOL 1-2, Listening). Students should take the ESOL ACP that is appropriate for the course that they are enrolled in. These tests will be administered in the Spring.	7-9	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	6-8	Same as #1
11. *Assessments of Course Performance*, year n, Language Arts (including ESOL, grades 10-12, first semester, grade 9, and first and second semester, grades 10-12), Mathematics, Social Studies, Science, Reading, World Language, (72 courses). Honors courses are considered separately. (raw score)	9-12	HLM on residuals (student level)	*Assessments of Course Performance*, year n-1, 72 courses, best predictors	7-11	Same as #1.
			Tests of Achievement and Proficiency, year n-1, Reading and Mathematics	9	
			Iowa Tests of Basic Skills, year n-1, Reading and Mathematics	7-8	
			Texas Assessment of Academic Skills, year n-1, Reading and Mathematics	7, 8, 10	
			Woodcock-Muñoz Language Survey, Broad Ability Score, year n-1	8-11	
12. *Woodcock-Muñoz Language Survey, Broad Ability Score*, year n, (raw score)	1-6	HLM on residuals (student level)	*Woodcock-Muñoz Language Survey, Broad Ability Score*, year n-1	K-11	Same as #1.
13. Graduation Rate, year n (percent graduated)	9-12	Multiple regression (school level)	Graduation Rate, years n-1 and n-2	9-12	None

BEST COPY AVAILABLE

14. *Scholastic Aptitude Tests* and *American College Test*, Percent Tested, year n (percent tested)	9-12 (statistic is based on percentage of twelfth graders who have ever taken the test)	Multiple regression (school level)	Percent Tested, *Scholastic Aptitude Test* and *American College Test*, years n-1 and n-2	9-12	None
15. *Scholastic Aptitude* Test and *American College Test*,, Verbal and Quantitative, year n (raw score)	9-12 (student's best score)	HLM on residuals (student level)	Various *Assessments of Course Performance*, year n-1, Language Arts, Mathematics, Social Studies, Science, Reading, World Language (72 courses)	8-11	Same as #1.
16. Dropout Rate, year n-1 (percent dropout)	7-12	Multiple Regression (school level)	Dropout Rate in years n-2 and n-3.	7-12	None
17. Student enrollment in Prehonors and Honors Courses, year n (percent in accelerated courses)	7-12	Multiple regression (school level)	Student enrollment in Prehonors and Honors Courses in years n-1 and n-2.	7-12	None
18. Student enrollment in Advanced Placement Courses, year n (percent in advanced diploma plans)	11-12	Multiple regression (school level)	Student enrollment in Advanced Placement Courses in years n-1 and n-2.	11-12	None
19. Percent Tested, *Preliminary Scholastic Aptitude Test*, year n (percent tested)	9-12	Multiple regression (school level)	Percent tested, *Preliminary Scholastic Aptitude Test*, years n-1 and n-2.	9-12	None

BEST COPY AVAILABLE

20. *Preliminary Scholastic Aptitude Test*, year n, Verbal and Quantitative (scale score)	9-12 (student's best score)	HLM on residuals (student level)	Various *Assessments of Course Performance*, year n-1, Language Arts, Mathematics, Social Studies, Science, Reading, World Language (72 courses)	8-11	Same as #1.
21. Percent of students passing the Advanced Placement Examinations, year n (denominator is number of students enrolled in Advanced Placement Courses)	11-12	Multiple Regression (school level)	Percent of students passing the Advanced Placement Examinations, year n-1 and n-2.	11-12	None

References

Aiken, L.S. and West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newburg Park: Sage.

Bano, S.M. (1985). *The Logic of Teacher Incentives*. Washington, D.C.: National Association of State Boards of Education.

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newburg Pass, California: Sage Publications.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon, R. (1988). *An Introduction to HLM: Computer Program User's Guide* (2nd ed.) Chicago, Ill: University of Chicago.

Cohen, J. (1968). *Multiple Regression As A General Data - Analytic Systems.* Psychological Bulletin, 70, 426-443.

Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences* (1st edition). Hillsdale, N. J. Lawrence Erlbaum. 2nd edition published in 1983.

Darlington, R. B. (1990) *Regression and Linear Models.* New York: McGraw-Hill

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.V., (1981). *Estimation in Covariance Components Models*. Journal of the American Statistical Association, 76, 341-353.

Elston, R.C. and Grizzle, J.E. (1962). *Estimation of Time Response Curves and Their Confidence Bands*. Biometrics, 18, 148-159.

Felter, M. and Carlson, D. (1985). *Identification of Exemplary Schools on a Large Scale.* In Austin and Gerber (eds.), Research on Exemplary Schools. New York: Academic Press, 83-96.

Goldstein, H., (1987). *Multilevel Models in Educational and Social Research,* New York: Oxford University Press.

Good, T.L., & Brophy, J.E. (1986). *School effects*. In M.C. Wittrock (Ed.), Handbook of Research on Teaching (3rd ed.) (pp. 570-602). New York, NY: Macmillan Publishing Company.

Haertel, E. (1986), *One Valid Use of Student Performance Measures for Teacher Evaluation.* Education Evaluation and Policy Analysis, 8, 45-60.

Henderson, C.R., (1984). *Applications of Linear Models in Animal Breeding*. Guelph, Canada: University of Guelph.

Iwanicki, E. G. (1986). *Continuities and Discontinuities Among Measures of Teacher Performance*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Jaeger, R. M. (1992). *Weak Measurement Serving Presumptive Policy*. Kappan, 74, 2, 118-128.

Kirst, M. (1986). *New Directions for State Education Data Systems*. Education and Urban Society, 18, 2, 343-357.

Klitgaard, R. E. and Hall, G. R. (1973). *A Statistical Search for Unustually Effective Schools*. Santa Monica, Ca.: Rand Corporation.

Laird, N.M. and Ware, H. (1982). *Random-Effects Models for Longitudinal Data*, Biometrics, 38-963-974.

Longford, N. T. (1987). *A Fast Scoring Algoritem for Maximum Likelihood Estimation In Unbalanced Mixed Models with Nested Random Effects* , Biometrika, 74 (4), 817-827.

McKenzie, D. (1983). *School Effectiveness Research: A Synthesis and Assessment*. In P. Duttweiler (ed.), Educational Productivity and School Effectiveness. Austin, Texas: Southwest Educational Development Laboratory.

Mason, W.M., Wong, G.Y. and Entwistle, B. (1984). *Contextual Analysis Through the Multilevel Linear Model*. In Leinhardt (ed.) Sociological Methodology, 1983-84, San Francisco: Josey-Bass, pp. 72-103.

Mendro, R.L. and Webster, W.J. (1993*). Using School Effectiveness Indices to Identify and Reward Effective Schools*. Paper presented at the Rocky Mountain Research Association, October, 1993, Las Cruces, New Mexico.

Millman, J. (ed.) (1981). *Handbook of Teacher Evaluation*. Beverly Hills, California, Sage.

Rosenburg, B. (1973). *Linear Regression With Randomly Dispersed Parameters*. Biometrika, 60, 61-75.

Saka, T. (1984). *Indicators of School Effectiveness: Which are the Most Valid and What Impacts Upon Them*? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA: (ERIC ED 306 277).

Sanders, W. L. (1997). *Private Conversation*.

Sanders, W. L., & Horn, S. P. (1995). *The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment,* in Shinkinlfred. A. J. and Stufflebeam, D. L., Teacher Evaluation: Guide to Effective Practice, Boston: Kluwer.

Webster, W. J. (1998). *A Comprehensive System for the Evaluation of Schools.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, California, April 13-17.

Webster, W. J. and Olson, G. H. (1988). *A Quantitative Procedure for the Identification of Effective Schools.* Journal of Experimental Education, 56, 213-219.

Webster, W. J. and Edwards, M. E. (1993). *An Accountability System for School Improvement.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, Georgia, April 12-16. ERIC TM 019 915.

Webster, W. J., Mendro, R. L., & Almaguer, T. D. (1993). *Effectiveness Indices: The Major Component of an Equitable Accountability System.* ERIC TM 019 193.

Webster, W.J., Mendro, R.L., and Almaguer, T. (1994). *Effectiveness Indices: A "Value Added" Approach to Measuring School Effect.* Studies in Educational Evaluation. 20, 113-145.

Webster, W. J., Mendro, R. L., Bembry, K. L., and Orsak, T. H. (1995). *Alternative Methodologies for Identifying Effective Schools.* A paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Webster, W. J., Mendro, R. L., Orsak, T. H., and Weerasinghe, D. (1996). *The Applicability of Selected Regression and Hierarchical Linear Models To The Estimation of School and Teacher Effects.*. A paper presented at the annual meeting of the American Educational Research Association, March, 1996, New York, NY.

Webster, W. J. and Mendro, R. L. (1997). *The Dallas Value-Added Accountability System,* in Grading Teachers, Grading Schools ,ed. by J. Millman, Newbury Park, CA, Sage , 81-99.

Webster, W. J., Mendro, R. L., Orsak, T., Weerasinghe, D., and Bembry, K. (1997). *Little Practical Difference and Pie in the Sky: A Response to Thum and Bryk and a Rejoinder to Sykes* in Grading Teachers, Grading Schools, ed. by J. Millman, Newbury Park, CA, Sage, 120-130.

Webster, W. J., Mendro, R L., Orsak, T. H., and Weerasinghe, D. (1997). *A Comparison of the Results Produced By Selected Regression and Hierarchical Linear Models in the Estimation of School and Teacher Effect,* paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois,

March 24-28. Accepted for publication in Multiple Linear Regression Viewpoints.

Weerasinghe, D., Orsak, T., and Mendro, R. (1997). *Value Added Productivity Indicators: A Statistical Comparison of the Pre-Test/Post-Test Model and Gain Model*. A paper presented at the 1997 Annual Meeting of the Southwest Educational Research Association, January, 1997, Austin, Tx.